Absolute pose estimation using multiple forms of correspondences from RGB-D frames

Shuda Li and Andrew Calway

Abstract— **We describe a new approach to absolute pose estimation from noisy and outlier contaminated matching point sets for RGB-D sensors. We show that by integrating multiple forms of correspondence based on 2-D and 3-D points and surface normals gives more precise, accurate and robust pose estimates. This is because it gives more constraints than using one form alone and increases the available measurements, especially when dealing with sparse matching sets. We demonstrate the approach by incorporating it within a RANSAC algorithm and introduce a novel direct least-square approach to calculate pose estimates. Results from experiments on synthetic and real data demonstrate improved performance over existing methods.**

I. Introduction

Absolute pose estimation from noisy and outlier contaminated matching point sets is a fundamental task when using RGB-D sensors for 3-D navigation and reconstruction [1], [2], [3]. Existing methods solve either the Perspective-*n*-Points (P*n*P) [4], [5], [6] or the Absolute Orientation (AO) [7], [8], [9] problem, using either 2-D to 3-D or 3-D to 3-D point correspondences, respectively. In this paper, we show that when both of these are combined and integrated with additional surface normal to surface normal correspondences, then the pose can be estimated with significantly higher accuracy, precision and robustness, as shown in Fig. 1a.

We formulate absolute pose estimation as follows (see Fig. 1b). Given a calibrated sensor, we want to estimate the rotation $\mathbf{R}$ and the camera centre $\mathbf{c}$ w.r.t a global map given a query RGB-D frame. The global map is composed of a point cloud and each point has a RGB feature descriptor, a 3-D location $\mathbf{q}_i$ and a surface normal $\mathbf{m}_i$. From the RGB-D frame, 2-D key points and RGB feature descriptors are extracted and each point has a 3-D position $\mathbf{p}_i$ and a surface normal $\mathbf{n}_i$ estimate. Feature descriptors are used to match with points in the global map and each pair of matched points provides up to 3 forms of correspondence, namely 2-D to 3-D (2-3), 3-D to 3-D (3-3) and surface normal to surface normal (N-N). The matches are assumed to contain outliers but to have sufficient inliers to allow pose estimation. Note that in contrast to motion tracking or relative pose estimation, for absolute pose estimation no prior camera pose is available.

Previous approaches use either the 2-3 or the 3-3 correspondences. Our approach is to use as many forms of correspondence as possible, inspired by the observation that the larger the number of independent measurements, the more reliable the estimation. Moreover, when matches are sparse due to poor image quality or lack of texture, using more forms of correspondence is the only option to increase measurements and improve pose estimation. It is known that a pair of 2-3 correspondences gives 2 non-linear constraints [10] and that a pair of 3-3 correspondences gives 3 linear constraints [8]. Similarly, a pair of N-N correspondences gives another 2 linear constraints to the rotation component of the pose. In all, allowing for dependencies between the 2-3 and 3-3 constraints, this gives 5 independent constraints when using all 3 forms of correspondence. Our experiments demonstrate that when these are integrated into a single algorithm it yields a significant improvement in pose estimation.

Our algorithm has two components: a RANSAC component to identify inliers; and a novel direct least-square component to estimate the pose from the inliers. In the former, pose candidates are generated by solving minimal sets depending on the availability of types of correspondence and each candidate is then voted on using all forms of correspondence. The flexibility offered by using different forms of correspondence in the candidate generating mechanism greatly increases the chance of sampling a non-degenerated minimal set and therefore reduces the number of RANSAC iterations needed to obtain a good pose estimate. In the least-square algorithm to solve for the pose, the rotation is obtained first by extracting all rotational constraints from the 2-3, 3-3 and N-N correspondences. This rotation is then used to estimate the camera center using constraints extracted from the 2-3 and 3-3 correspondences (the N-N correspondences have no impact on camera centre). In both cases, dynamic weighting of the correspondence terms is used to take account of the uncertainty in the 2-D and 3-D position and normal estimates. Results from synthetic and real data experiments demonstrate that both the RANSAC method alone and RANSAC followed by least-square optimisation give significantly more robust, accurate and precise estimates than state-of-the-art solutions using single forms of correspondence [9], [6]. Source code of our implementation is available at `github.com/Mavericklsd/rgbd_pose_estimation`.

II. Related work

Previous approaches to pose estimation solve either the AO problem [7], [8], [9] or the P*n*P problem [4], [5]. The former aligns one set of 3-D points with another and the minimal solvable case is 3 pairs of 3-3 correspondences, with each pair providing 3 linear constraints (non-colinear). The first direct solution for the minimal and over-determined sets was introduced by Arun *et al.* [8]. Independently, Horn *et al.* [7] developed a similar approach using quaternions. Later, Shinji [9] proved that the approach of Arun *et al.* is

The authors are with the Department of Computer Science, University of Bristol, {csxsl, csadc}@bristol.ac.uk.

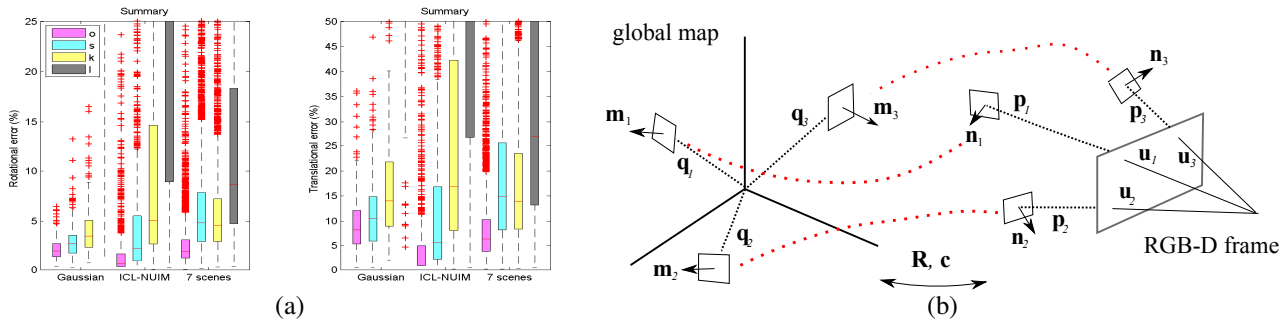


Fig. 1: (a) Rotation and translation errors for synthetic (Gaussian and ICL-NUIM) and real world (7 scenes) data, demonstrating that the proposed method 'o' (RANSAC with optimisation and dynamic weighting) significantly out-performs state-of-the-art methods 's' [9], 'k' [6] and 'l' [4]; (b) Pose estimation using multiple forms of correspondence 2-3 ($\mathbf{u}_i \leftrightarrow \mathbf{q}_i$), 3-3 ($\mathbf{p}_i \leftrightarrow \mathbf{q}_i$) and N-N ($\mathbf{n}_i \leftrightarrow \mathbf{m}_i$).

equivalent to a least squares optimisation and also solved the reflection problem in these previous methods.

The PnP problem is to determine pose from 2-3 correspondences. In this case, each pair provides 2 independent nonlinear constraints and the minimal solvable case is P4P: 3 pairs of non-colinear 2-3 correspondences contain 6 independent non-linear constraints which yields 4 solutions and the extra pair is required to determine the valid one [11], [6]. The method described by Kneip *et al.*. [6] represents the state-of-the-art solution to the P4P problem with high numerical stability and low possibility of being affected by degenerate configurations.

For $n > 4$, the PnP problem has been extensively studied. Recent methods include the non-iterative approach described by Lepetit *et al.* [4] which provides an $O(n)$ solution, contrasting with the $O(n^5)$ of state-of-the-art methods at the time. The method was further improved upon in [10] and [5], which dealt with inaccuracies when n is small. Comprehensive surveys of other methods can be found in [4] and [5]. However, none of these approaches make use of constraints from depth and all of them involve solving complex non-linear equations and having to deal with multiple solutions.

III. Three forms of correspondence

In this section we define each of the 3 forms of correspondence. For 2-3 correspondence, we have a set of 3-D global map points defined in a world coordinate system W, $\mathbf{q}_i \in \mathbb{R}^3$, and their corresponding projections onto the RGB-D image plane, $\mathbf{u}_i$. The $\mathbf{u}_i$ is the homogeneous form coordinate of corresponding 2-D key points on the image. For a sensor with pose defined by a rotation matrix $\mathbf{R} \in \mathbb{SO}_3$ and the camera centre $\mathbf{c} \in \mathbb{R}^3$ w.r.t W, then assuming the sensor is fully calibrated, each 2-3 correspondence should satisfy $\lfloor \mathbf{u}_i \rfloor = \mathbf{R} \lfloor \mathbf{q}_i - \mathbf{c} \rfloor$, where $\lfloor . \rfloor$ denotes the unit normalization operator. For 3-3 correspondence, we have two sets of 3-D points $\mathbf{p}_i \in \mathbb{R}^3$ and $\mathbf{q}_i \in \mathbb{R}^3$, where the $\mathbf{p}_i$ are defined in the sensor coordinate system C and the $\mathbf{q}_i$ is in the world coordinate system W. In this case each pair of correspondences should satisfy $\mathbf{p}_i = \mathbf{R}(\mathbf{q}_i - \mathbf{c})$. Finally, we define a surface normal to surface normal (N-N) correspondence as a pair of surface normals which are associated with a pair of 3-3 correspondences. We denote a pair of N-N correspondences as $\mathbf{n}_i$, $\mathbf{m}_i$ which are both unit vectors in $\mathbb{R}^3$ and are defined in the sensor C and world W coordinate systems, respectively. The relationship between the pair is given by the rotation component of the sensor pose (independent of the camera centre), i.e. each pair should satisfy $\mathbf{n}_i = \mathbf{R}\mathbf{m}_i$.

Given the above relationships, we can combine all 3 forms of correspondence into a single error function as a means for determining least-square estimates of $\mathbf{R}$ and $\mathbf{c}$. This can formulated as follows

$$\begin{aligned} e^2(\mathbf{R},\mathbf{c}) = & \frac{\psi}{|\Lambda_1|} \sum_{i \in \Lambda_1} w_i \left\| \lfloor \mathbf{u}_i \rfloor - \mathbf{R} \lfloor \mathbf{q}_i - \mathbf{c} \rfloor \right\|^2 + \\ & \frac{1}{|\Lambda_2|} \sum_{i \in \Lambda_2} v_i \left\| \mathbf{p}_i - \mathbf{R}(\mathbf{q}_i - \mathbf{c}) \right\|^2 + \\ & \frac{\psi}{|\Lambda_3|} \sum_{i \in \Lambda_3} \lambda_i \left\| \mathbf{n}_i - \mathbf{R}\mathbf{m}_i \right\|^2 \end{aligned} \tag{1}$$

where ψ is a weight balancing the relative contribution of the different forms of correspondence and Λ_1, Λ_2 and Λ_3 denote the sets of matching pairs with 2-3, 3-3 and N-N correspondences, respectively. $|\Lambda_i|$ is the number of elements in the i-th set. The first and third terms are given the same weight in this equation since they both involve minimising unit vectors. In addition, individual correspondence pairs are weighted dynamically using w_i, v_i and λ_i according to the certainty of the associated 2-D position, 3-D position and surface normal estimates. These are defined in Section VI.

IV. RANSAC using all correspondences

We now describe a RANSAC algorithm to minimise (1) given a set of noisy matched pairs contaminated with outliers. Each pair will contribute either 1 (2-3), 2 (2-3 and 3-3) or 3 (2-3, 3-3 and N-N) forms of correspondence. We first randomly sample minimal sets of pairs from which we can generate candidate poses. We choose minimal sets to give a mix of the 3 forms of correspondence. Support for each potential pose is then sought from the complete set, with votes derived from correspondences based on the respective terms in (1). The pose with the highest number of votes is then selected and its inliers are used to obtain a pose estimate using the least-square optimisation in Section V.

There are a number of possibilities for choosing minimal sets given that the 3 forms of correspondence provide 5 independent constraints. We have opted to use those for which solutions are available and to do so in a manner which aims to provide a good mix of the 3 forms. These are: 2 pairs, both with 3-3 correspondences and one with N-N correspondence; 3 pairs, each with 3-3 correspondence; and 4 pairs, each with 2-3 correspondence. To solve the first we adapt the algorithm described by Drost *et al*.[12] and the other two are solved using the algorithms in [9] and [6] .

The minimal sets are used as follows. In each RANSAC iteration, we randomly select 4 matching pairs. Within these, we identify at most one of each minimal set, choosing randomly if more than one is present (recall that each matching pair can have between 1 and 3 forms of correspondence). Each selected set is then solved, giving between 1 and 3 candidate poses. Note that every pair will provide a 2-3 correspondence and hence 4 pairs will generate at least 1 candidate pose. We found that this approach means that we have a good chance of utilising a mix of correspondences, increasing the chances of sampling valid minimal sets and reducing the number of RANSAC iterations.

Support for each candidate pose is then evaluated using the remaining pairs in the matching set. For each pair, each form of correspondence (if present) votes for the pose if the corresponding error term in (1) is within a threshold. Thus each pair can contribute up to 3 votes for a given pose candidate. For example, given a pose candidate $(\mathbf{R},\mathbf{t})$, the ith pair having only 2-3 correspondence, say, i.e. lacking depth information due to reflections in the scene, for example, would contribute 1 vote to the candidate if $\|\lfloor\mathbf{u}_i\rfloor - \mathbf{R}\lfloor\mathbf{q}_i - \mathbf{c}\rfloor\|^2 < \tau$ for an outlier threshold τ, where $\mathbf{u}_i$ and $\mathbf{q}_i$ are its 2-D to 3-D corresponding points. Similarly, a pair having 3 forms of correspondence would test each of them against the relevant terms in (1) and vote accordingly.

Following a fixed number of iterations, the candidate pose achieving the most votes is then selected. We found that the resulting pose estimates were good and invariably better than those obtained using single forms of correspondence. However we also found that optimising a pose estimate from the inliers of the selected pose yielded further improvement. In the next section we describe how this can be done using a novel direct approach.

V. Direct least-square solution

We seek to minimise (1) to obtain estimates for $\mathbf{R}$ and $\mathbf{c}$ given a set of matching pairs and their associated correspondences. These are assumed to be inliers resulting from the RANSAC process. We use a novel non-iterative algorithm based on the direct least-square optimisation of 3-3 correspondence solution proposed by Shinji [9]. For clarity, we summarise the three stages of that algorithm given a set of N 3-D corresponding points $\mathbf{p}_i$ and $\mathbf{q}_i$:

1) Remove the translation by subtracting the centroid from each set of points: $\mathbf{q}'_i = \mathbf{q}_i - \bar{\mathbf{q}}$ and $\mathbf{p}'_i = \mathbf{p}_i - \bar{\mathbf{p}}$.
2) If $\mathbf{UDV}^T$ is the singular value decomposition (SVD) of covariance matrix $\Sigma = \frac{1}{N}\sum \mathbf{p}'_i \mathbf{q}'^T_i$, then the optimized rotation estimate is given by $\widehat{\mathbf{R}} = \mathbf{USV}^T$, where $\mathbf{S}$ is either the identity matrix $\mathbf{I}$ or $\mathbf{I}$ with its last diagonal element replaced by -1 if $det(\Sigma) < 0$.
3) The optimized camera centre in the world coordinate system can be estimated using $\widehat{\mathbf{c}} = \bar{\mathbf{q}} - \mathbf{R}^T\bar{\mathbf{p}}$.

We extend the above algorithm to include all correspondences available. The N-N correspondences are independent of the camera centre and hence can be straightforwardly combined with the covariance matrix Σ. Given that the camera centre is available from either the method in [9] using 3-3 correspondences, the method in [13], or simply from the output of the RANSAC algorithm, the rotational constraints from 2-3 correspondences can be extracted from (**??**) and can be combined with the covariance matrix in the same way as the N-N correspondences. Overall, the extended covariance matrix is:

$$\begin{aligned}\Sigma = &\frac{\psi}{|\Lambda_1|}\sum_{i\in\Lambda_1} w_i \lfloor\mathbf{u}_i\rfloor \lfloor\mathbf{q}_i - \mathbf{c}\rfloor^T + \\ &\frac{1}{|\Lambda_2|}\sum_{i\in\Lambda_2} v_i \mathbf{p}'_i \mathbf{q}'^T_i + \frac{\psi}{|\Lambda_3|}\sum_{i\in\Lambda_3}\lambda_i \mathbf{n}_i \mathbf{m}_i^T\end{aligned} \tag{2}$$

Since we want both the N-N and 3-3 correspondences to play equal roles in determining the rotation we set $\psi = \frac{1}{|\Lambda_2|}\sum|\mathbf{p}'_i|^2$. Another option is to normalize $\mathbf{p}'_i$ and $\mathbf{q}'_i$ to make them unit vectors. However, we found that this decreases the precision of the estimated rotation matrix since the length of $\mathbf{p}'_i$ and $\mathbf{q}'_i$ are lost after normalization.

After obtaining the optimized $\widehat{\mathbf{R}}$ via SVD of Σ as in step 2 above, we then optimize the camera centre $\widehat{\mathbf{c}}$ in W. Shinji's algorithm can estimate $\widehat{\mathbf{c}}_s$ by optimizing 3-3 but not 2-3 correspondences. Below, we derive a linear solution to make use of 2-3 correspondences. The resulting $\widehat{\mathbf{c}}'$ is then combined with the $\widehat{\mathbf{c}}_s$ from Shinji's algorithm.

Inspired by the Slabaugh *et al*.. [14], we found that the camera centre in the world coordinate system can be retrieved by finding the optimal intersection of visual rays. The direction of visual rays in world coordinates can be calculated by rotating the visual ray from the senor coordinate $\mathbf{u}'_i = \widehat{\mathbf{R}}^T\lfloor\mathbf{u}_i\rfloor$. The line segment from a $\mathbf{q}_i$ to the camera centre $\widehat{\mathbf{c}}'$ should be identical to the projection of the line segment onto the visual ray $\mathbf{u}'_i$:

$$\sum_{i=1}^{K} w_i\left[(\mathbf{q}_i - \widehat{\mathbf{c}}') - \mathbf{u}'_i\mathbf{u}'^T_i(\mathbf{q}_i - \widehat{\mathbf{c}}')\right] = 0 \tag{3}$$

Denoting $\widehat{\mathbf{c}}' = [x,y,z]^T$, $\mathbf{q}_i = [x_i,y_i,z_i]^T$ and $\mathbf{u}'_i = [a_i,b_i,c_i]^T$, the above equation can be expanded as

$$\begin{aligned}
&\sum_{i=1}^{K} w_i\{(x_i - x) - a_i[a_i(x_i - x) + b_i(y_i - y) + c_i(z_i - z)]\} = 0\\
&\sum_{i=1}^{K} w_i\{(y_i - y) - b_i[a_i(x_i - y) + b_i(y_i - y) + c_i(z_i - z)]\} = 0\\
&\sum_{i=1}^{K} w_i\{(z_i - z) - c_i[a_i(x_i - y) + b_i(y_i - y) + c_i(z_i - z)]\} = 0
\end{aligned} \tag{4}$$

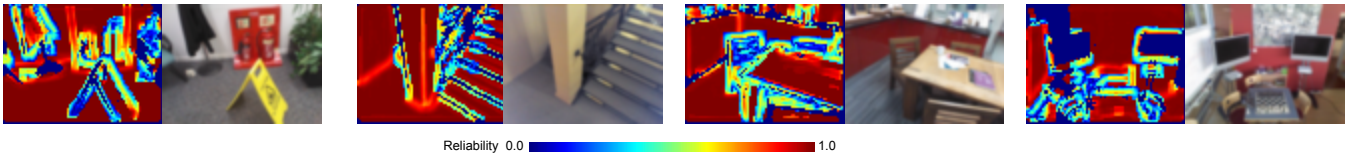


Fig. 2: Dynamic weights for N-N correspondences. The color maps show the distribution of the weights. They are calculated using 5x5 patches over 80x60 depth maps. Note that the plane of the monitors are assigned as 0 in the right sub-figure. This is because there are no depth values. The RGB-D frames are from the '7 scenes' dataset.

which can be written as $\sum_{i=1}^{K} w_i \mathbf{A}_i \widehat{\mathbf{c}}' = \sum_{i=1}^{K} w_i \mathbf{A}_i \mathbf{q}_i$, where

$$\mathbf{A}_i = \begin{bmatrix} 1-a_i^2 & -a_i b_i & -a_i c_i \\ -a_i b_i & 1-b_i^2 & -b_i c_i \\ -a_i c_i & -b_i c_i & 1-c_i^2 \end{bmatrix} \quad (5)$$

from which we obtain $\widehat{\mathbf{c}}' = \mathbf{A}^{-1}\mathbf{b}$, where $\mathbf{A} = \sum_{i=1}^{K} w_i \mathbf{A}_i$ and $\mathbf{b} = \sum_{i=1}^{K} w_i \mathbf{A}_i \mathbf{q}_i$. Finally, the camera centre is calculated by combining $\widehat{\mathbf{c}}_s$ and $\widehat{\mathbf{c}}'$, i.e. $\widehat{\mathbf{c}} = K\widehat{\mathbf{c}}'/(K+N) + N\widehat{\mathbf{c}}_s/(K+N)$

VI. Dynamic weights

Ideally, in (1), we want each pair of correspondences multiplied with a dynamic weight (w_i, v_i or λ_i) such that if a correspondence is more likely to be corrupted, it will be assigned a small weight and vise versa.

To weight 2-3 correspondences, we adopt the common practice of $w_i = 1 - \frac{d_{1st}}{d_{2nd}}$, where d_{1st} is the feature distance score between a 2-D key point and its first nearest neighbour (1-NN) in the global map and d_{2nd} is the score to 2-NN. The ratio represents the reliability of the match; the smaller the ratio, the more reliable the match. To weight 3-3 correspondences, we use the method introduced by Nguyen et al. [15] to take account of errors within the RGB-D sensor.

To weight N-N correspondences, we introduce a novel approach. Motivated by the observation that surface normals are less reliable around corners, edges and occlusion boundaries and are very stable inside planar regions, we weight the N-N correspondences according to a principle component analysis (PCA) over a fixed patch. This has the advantage that weight calculation comes with surface normal extraction at almost no extra cost [16], [17].

Specifically, we calculate the eigenvalues and eigenvectors of the covariance matrix over the patch in the depth image used to compute the normal. The eigenvector corresponding to the smallest eigenvalue gives the direction of the normal. The dynamic weight for the N-N correspondence is then obtained from the 2nd and 3rd eigenvalues E_2 and E_3: $\lambda = (E_2 - E_3)/(E_2 + E_3)$. Essentially, $\lambda \in [0,1]$ measures the surface curvature and roughness of the patch. Fig. 2 shows examples of weights computed for RGB-D frames. Note that areas around high curvature and occlusion boundary are correctly assigned with low weights indicating low reliability.

VII. Experiments

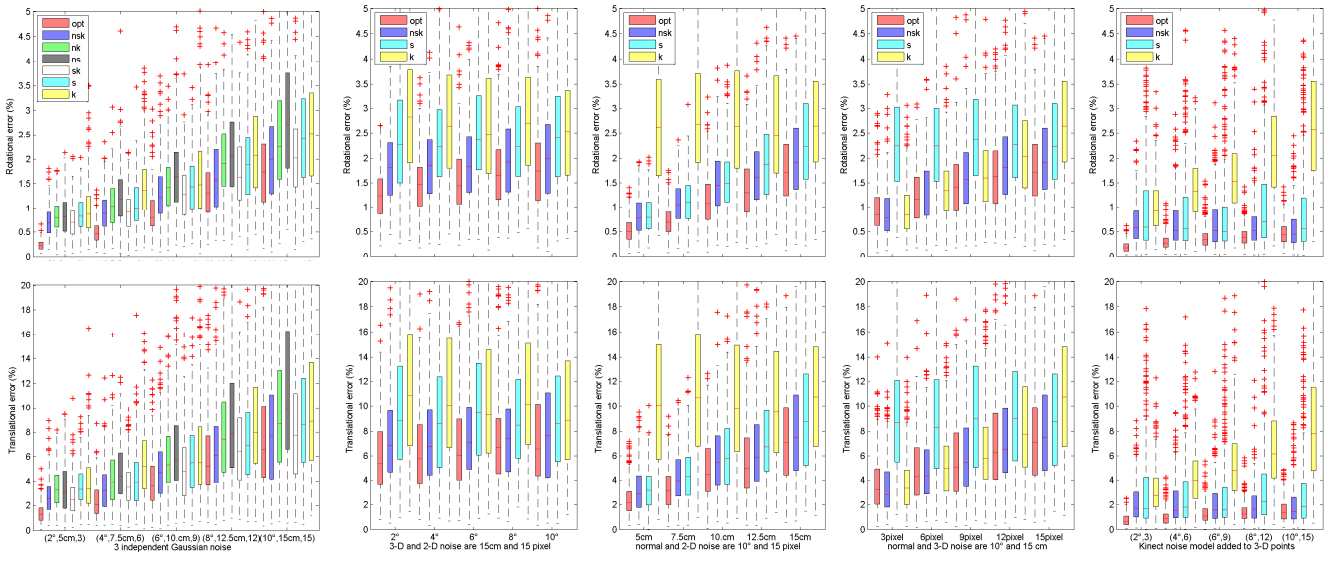
We evaluated the above algorithms using both synthetic and real data sets. For comparison, we implemented RANSAC-based pose estimation based on Kneip *et al.*'s P3P [6] and Shinji's AO [9] algorithms, which are referred to as 'k' and 's' respectively. Error metrics for rotation and camera centre were adopted from those used in [4]. The translational error is the distance from the estimated camera centre to the ground truth camera centre. We denote our three algorithms by 'nsk' (RANSAC only), 'opt' (RANSAC with direct least-square optimisation) and 'dw' (RANSAC with least-square optimisation and dynamic weighting).

In the first 2 synthetic experiments, we generated a set of 100 sparse points uniformly distributed within a viewing frustum (minimum distance 0.4m, maximum distance 8.m). They were then transformed to world coordinates using the ground truth rotation and camera centre and projected onto a virtual 640x480 image plane with focal length 585 and principle point at the image centre. The normals were uniformly distributed and facing toward the image plane. We carried out separate experiments using added Gaussian noise and Kinect noise [15]. 2-D and 3-D outliers are randomly selected within the image and the viewing frustum, respectively.

The algorithms were tested with various levels of noise added to each of the 3 forms of correspondence. First, we varied the noise from low to high simultaneously across all 3 and then fixed 2 at a high noise level whilst varying the 3rd from low to high. Each experiment was run 300 times. The outlier rejection threshold was set at 8 pixels, 6^o and 20cm, respectively, for the 2-3, 3-3 and N-N correspondences. RANSAC iterations were set as 200 and outlier ratio as 0.5 to generate all plots except Fig. 4. By way of comparison, we also investigated using combinations of 2 forms of correspondence: N-N and 3-3, denoted 'ns'; N-N and 2-3, denoted 'nk'; 3-3 and 2-3, denoted 'sk'.

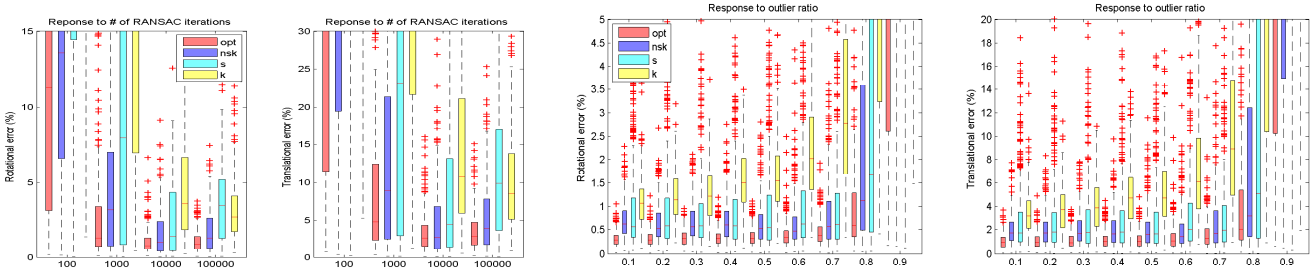
The results are illustrated using a standard box plot showing the mean error and the precision. In Fig. 3, we can see that when using all forms of correspondence, the pose estimation is consistently more accurate and more precise than using 3-3 or 2-3 correspondences alone. Using 2 are better than 1 with the exception of 'ns' and 'nk'. They represent the combination of N-N and 3-3 and N-N and 2-3, respectively. Fig. 4 shows the performance as the number of RANSAC iterations and the outlier ratios are varied.

RGB-D sensors are known to have unique noise characteristics and we investigated the effects of using a more realistic noise model for the depth. We used that in [15], in which the depth noise increases with distance from the sensor. The angle between the visual ray and the principle axis and the angle between the surface normal and the principle axis are proportional to the 3-D noise. Results for different noise levels are shown in Fig. 3e. Parameter settings were identical to that used in the other experiments and the results again



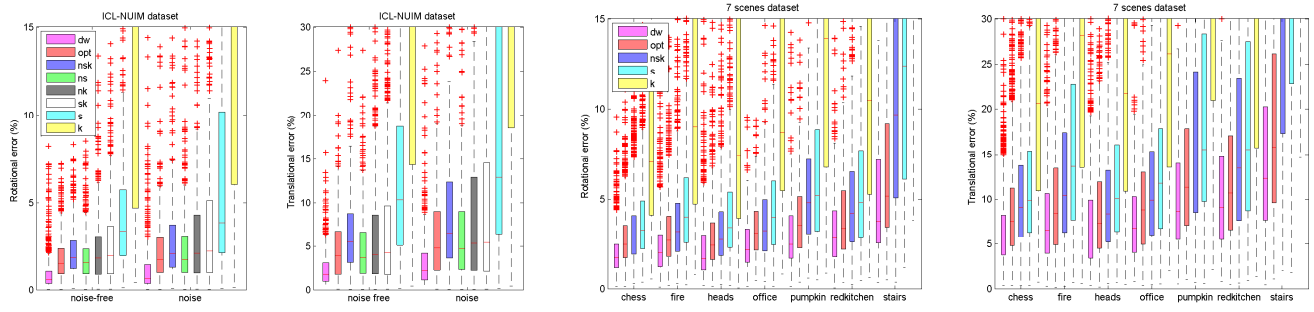
(a) (b) (c) (d) (e)

Fig. 3: Comparing the accuracy of our method against baseline approaches. Top and bottom rows corresponds to rotational and translational error. 30 points with normal attached are used. The outlier ratio is 0.25. In column (a) We varied the 3 noise sources simultaneously; in (b) we kept the 2-D and 3-D noise at high level and varied the normal noise; in (c) we kept 2-D and normal noise at high level and varied 3-D noise; in (d) 3-D and normal noise were at high level and 2-D noise was varied; in (e) 3-D noise was based on simulating Kinect-like sensor noise.



(a) (b) (c) (d)

Fig. 4: (a) & (b) comparison of our approaches with the baseline using different RANSAC iterations; (c) & (d) response of the proposed method to outlier ratio. The standard deviation for the correspondence noise for 2-3, 3-3 and N-N was 9 pixel, 7.5cm and 6^o, respectively.



(a) (b) (c) (d)

Fig. 5: Comparing the accuracy of our method against baseline approaches using public RGBD datasets. The left two figures are results from using ICL-NUIM dataset. The right two figures show corresponding results from using the '7 scenes' dataset.

show the advantage of the proposed method.

We also evaluated the algorithms using the ICL-NUIM dataset [18] consisting of RGB-D frames and ground-truths obtained by ray-tracing 3-D models with realistic image and depth noise. We integrated our pose estimation into our RGB-D mapping and relocalisation framework as described in [3]. The global map was created using the 3rd sequence from the 'living room' and our approach was tested by relocalising frames from the remaining 3 sequences. The BRISK feature descriptor [19] was used to match with points in global

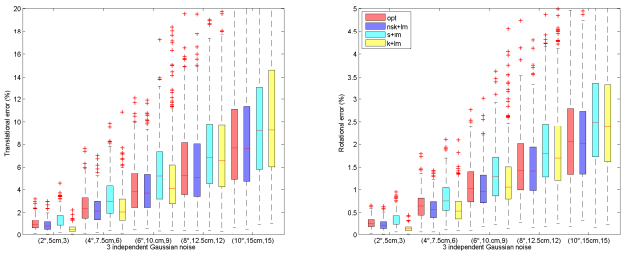


Fig. 6: Comparing the accuracy of our method against iterative non-linear optimisation (LM) using synthetic dataset. The proposed direct approach (red) gives equivalent performance to non-linear optimisation using 3 forms of correspondences (blue) and both are more accurate than using single forms of correspondence, especially when the data is noisy.

map. The parameter settings were identical to that used in the other experiments. We compared performance using 3 and 2 forms of correspondence against using a single form of correspondence. The results are shown in Fig. 5a-b and they confirm the trend shown earlier: using all 3 forms of correspondences out-performs using 2 or 1 and using 2 are better than 1 with the exception of using N-N and 3-3.

We also tested the method on the real RGB-D dataset used in [20] composed of RGB-D frames captured from 7 different scenes. We built global maps using the‘’training’ sequences and use our RGB-D pose estimation method to relocalise the testing sequences. The results are shown in Fig. 5c and 5d and again confirm that superior performance of our method.

Finally, we also compared the accuracy of the direct least-square approach with that obtained using an iterative non-linear optimisation. We used Levenberg-Marquardt (LM) to optimise (4) from the inlier set resulting from RANSAC and the results are shown as the blue bars, ’nsk+lm’, in Fig. 6. The yellow bars, ’k+lm’ and the cyan bars ’s+lm’ are the results of using LM to optimize the first and second terms, respectively, i.e. the errors in 2-3 and 3-3 correspondences. The results of using the direct least-square optimisation using all 3 forms of correspondences are shown in red, ’opt’, and are consistently better than using single forms of correspondences whilst delivering comparable results to that of LM using all 3. This demonstrates that the direct approach gives comparable accuracy whilst being computationally more efficient than non-linear optimisation by avoiding the need for iteration.

VIII. Conclusion

We have presented a novel pose estimation method for RGB-D sensors which integrates 2-D to 3-D, 3-D to 3-D and normal to normal correspondences. Results from experiments on synthetic and real data show that the method gives more accurate, more precise and more robust estimates than existing methods based on single forms of correspondence. Future work will investigate using alternative minimal sets and using the method on dynamic scenes.

References

[1] S. Lieberknecht, A. Huber, S. Ilic, and S. Benhimane, “RGB-D camera-based parallel tracking and meshing,” in *IEEE/ACM Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.

[2] J. Martinez-Carranza, A. Calway, and W. Mayol-cuevas, “Enhancing 6D Visual Relocalisation with Depth Cameras,” in *Intl. Conf. on Intelligent Robot Systems (IROS)*, 2013.

[3] S. Li and A. Calway, “RGBD Relocalisation Using Pairwise Geometry and Concise Key Point Sets,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015.

[4] V. Lepetit, F. Moreno-Noguer, and P. Fua, “EPnP: An Accurate O(n) Solution to the PnP Problem,” *Intl. Journal of Computer Vision (IJCV)*, vol. 81, no. 2, 2009.

[5] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, and M. Okutomi, “Revisiting the PnP Problem: A Fast, General and Optimal Solution,” in *Intl. Conf. on Computer Vision (ICCV)*, 2013.

[6] L. Kneip, D. Scaramuzza, and R. Siegwart, “A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation,” in *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[7] B. K. P. Horn, “Closed-form solution of absolute orientation using unit quaternions,” *Journal of the Optical Society of America A*, vol. 6, no. 4, 1987.

[8] S. K. Arun, T. S. Huang, and S. D. Blostein, “Least-Squares Fitting of Two 3-D Point Sets,” *IEEE Trans. Pattern Anal. Machine Intell. (PAMI)*, 1987.

[9] U. Shinji, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Trans. Pattern Anal. Machine Intell. (PAMI)*, 1991.

[10] S. Li, C. Xu, and M. Xie, “A Robust O(n) Solution to the Perspective-n-Point Problem,” *IEEE Trans. Pattern Anal. Machine Intell. (PAMI)*, vol. 34, no. 7, 2012.

[11] X.-s. Gao, X.-r. Hou, J. Tang, and H.-f. Cheng, “Complete solution classification for the perspective-three-point problem,” *IEEE Trans. Pattern Anal. Machine Intell. (PAMI)*, vol. 25, no. 8, 2003.

[12] B. Drost, M. Ulrich, N. Navab, and S. Ilic, “Model globally, match locally: Efficient and robust 3D object recognition,” in *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[13] J. a. Hesch and S. I. Roumeliotis, “A Direct Least-Squares (DLS) method for PnP,” in *Intl. Conf. on Computer Vision (ICCV)*, 2011.

[14] G. Slabaugh, R. Schafer, and M. Livingston, “Optimal Ray Intersection For Computing 3D Points From N -View Correspondences,” Georgia Tech, Tech. Rep., 2001.

[15] C. V. Nguyen, S. Izadi, and D. Lovell, “Modeling kinect sensor noise for improved 3D reconstruction and tracking,” in *3D Imaging Modeling Processing Visualization Transmission (3IMPVT)*, 2012.

[16] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, “Comparison of surface normal estimation methods for range sensing applications,” in *ICRA*, 2009.

[17] R. B. Rusu and S. Cousins, “3D is here : Point Cloud Library (PCL),” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2011.

[18] A. Handa, T. Whelan, J. Mcdonald, and A. J. Davison, “A Benchmark for RGB-D Visual Odometry , 3D Reconstruction and SLAM,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.

[19] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary Robust invariant scalable keypoints,” in *Intl. Conf. on Computer Vision (ICCV)*, 2011.

[20] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images,” in *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.